International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Modeling Cholera Outbreaks in Kenya Using Machine Learning Algorithms

Paul Njuguna Theuri

ABSTRACT

Kenya continues to grapple with the cholera outbreak due to environmental, socio-economic and behavioral factors. The traditional surveillance systems are largely used for reactive surveillance, which limits the capacity to foresee outbreaks. This study examines how machine learning can be used to forecast cholera outbreaks based on environmental, epidemiological and socio-economic inputs. Variables such as weekly cholera case counts, lagged trends in cholera case counts, and rainfall, temperature and sanitation variables were used to train a Random Forest classification model. Data preprocessing included data cleaning, encoding, and feature engineering, such as creating lag variables to represent temporal features. The data set was divided into a training and test set (80:20). The Synthetic Minority Oversampling Technique (SMOTE) was used to solve the class imbalance problem between the outbreak and non-outbreak events. The results indicated that there were temporal patterns of cholera transmission, with the most important factors were the presence of temporal features especially lagged case counts. The overall accuracy of the baseline model was good, but sensitivity to identify outbreak cases is low. The model showed good outcomes in detecting outbreaks, but with a slight decrease in overall accuracy once SMOTE was used, illustrating a trade-off between precision and recall in imbalanced data sets. Other environmental factors (temperature and rainfall) were also factors in prediction performance. The authors find machine learning models have the potential to predict cholera outbreaks, but the data is limited and has class imbalance and features that are not well represented. The next round of research needs to link multiple data and sophisticated scientific models in real time to improve prediction capabilities and provide the tools for early warnings.

CHAPTER 1:
INTRODUCTION

1.1 Background

Cholera is a diarrheal disease caused by the bacterium, Vibrio cholerae; the primary mode of transmission is through contaminated food and water. There are two main serogroups O1 and O139 that are known as the most responsible for major epidemics of cholera worldwide (Dossou Sodjinou et al., 2022). Cholera is highly widespread in those regions with ineffective sanitation and low availability of safe drinking water, which disproportionally concerns developing nations, such as Kenya World Health Organization (WHO, 2023).

Vibrio cholerae enters via contaminated water or food and is able to survive the acidic stomach environment, colonizes the small intestine with colonization factors and is able to proliferate in those cells. After colonization, the bacteria express a cholera toxin that is an enterotoxin of the AB_5 type, which activates adenylate cyclase to increase intracellular cyclic AMP, causing profuse electrolyte and fluid secretion. The clinical signs include excessive watery diarrhoea (typically 'rice-water stool'), vomiting and

muscle cramps, which, if untreated, can cause rapid dehydration and be fatal (Frontiers in Microbiology, 2023).

Meteorological information shows that cholera is a continuous menace in Kenya and has been reported to occur regularly during rainy seasons when flooding makes water sources contaminated. In Kenya, for instance, public health measures have had no impact on the continued vulnerability, as the Ministry of Health has reported on recurrent outbreaks in counties (Kenya Ministry of Health, 2023). Moreover, the transmission risk of cholera is much increased in informal settlements due to overcrowding, insufficient basic water and sanitation services, and urbanization. In informal settlements, the rainy seasons are important times for the spread of cholera because of the increased risk of flooding and contamination of water sources. In recent instances, for example, outbreaks have been reported in seven counties including Nairobi, Mombasa, Kisumu, Migori, Homa Bay, Kwale and Turkana, leading to many deaths in the wake of heavy rains. (Daily Nation 2025, June 18)

Stool culture and rapid diagnostic tests (RDTs) are essential tools for early detection of the cholera. However, access to these services is not as widespread in rural areas as in urban areas, and that slows down reports and reduces response strategies. Logistics problems, the low awareness of the public about cholera prevention and control measures (WHO 2023), are among the challenges faced by treatments such as oral rehydration solutions (ORS) and antibiotics in situations of resource limitations.

1.2 Problem Statement

Although cholera has been and remains a serious public health problem in Kenya, current prediction models rely primarily on past epidemiology data and simple statistical methods. The models do not usually consider the complex interactions between environmental factors and socio-economic factors that affect the outbreak of cholera. Furthermore, while there is some research on cholera prediction using machine learning in Kenya, existing studies are limited and do not provide a thorough relative analysis of numerous machine learning algorithms. This study aims to overcome such gap by developing a comprehensive and effective model that incorporates both socio-economic and environmental indicators to predict cholera. The purpose of this study is to assess and compare various machine learning algorithms to see which one is most effective at predicting cholera outbreaks in Kenya, thereby potentially providing better early warning systems and public health response capabilities.

1.3 Aim and Objectives

1.3.1 Main objective

To evaluate the effectiveness of the Random Forest machine learning model in predicting cholera outbreaks in Kenya.

1.3.2 Specific Objectives:

- Examine key environmental and socio-economic factors that influence the outbreak of cholera in Kenya.
- To create a trained Random Forest model and assess the model's performance on standard metrics for cholera outbreak prediction.
- To evaluate the effect of handling techniques for class imbalance (e.g. SMOTE) on model performance.

1.4 Research Questions

1. What are the effects of environment and socio-economic factors on cholera outbreaks in Kenya?

2. What is the accuracy with which the Random Forest algorithm can forecast cholera outbreaks in Kenya?
3. What effect does class imbalance handling technique, SMOTE, have on the performance of the Random Forest model?

1.5 Justification

This study is important as the researchers will have a comparative study of the various machine learning techniques to be used to ensure that the best model will be used to serve the predicted need for dealing with an outbreak in Kenya. This study will use environmental and socio-economic information to help inform early warning systems, which will improve cholera response strategies. Additionally, the study will test if multiple algorithms are applicable and able to be used in cholera outbreak prediction, which may further enhance prediction accuracy.

1.6 Significance and Anticipated Output

This work is hoped to contribute to the understanding of how cholera works in Kenya to more effectively integrate it into public health strategies. Cholera is a critical public health priority and outbreak threat, especially in resource-scarce countries, generally endemic to this infectious disease. The machine learning approach towards outbreaks proposed here aims to address a significant gap in the existing knowledge and will offer useful inputs to the health authorities.

1.7 Scope and limitations

Scope:

Geographical Scope: Target areas in Kenya that are prone to cholera, such as the interior of Nairobi slums, coastal zones and lake areas.

Data Scope: Use environmental, socio-economic and epidemiological information to develop a predictive model.

Gives mathematical knowledge in using Random Forest machine learning algorithm for model development and evaluation.

Limitations:

Limited data availability: Not having access to the real-time or full data in remote areas.

External Factors: Unpredictable variables could impact prediction accuracy.

Generalizability: It is possible that the model will not be easily transferable to other areas without modifications.

CHAPTER 2:LITERATURE REVIEW

2.1 Theoretical framework

This study is informed by four theories that are woven together to offer a solid basis for understanding and modelling a cholera outbreak through machine learning. These include:

2.1.1 Epidemiological Triad Theory (Primary Theory)

This is the main theoretical framework underpinning this study. It defines the cause of disease, in terms of interaction among three factors: Vibrio cholerae, human population, and environment (sanitation, water quality, etc.) (Park, 2009).

This theory directly guides the identification and incorporation of the epidemiological, environmental, and socio-economic factors into the machine learning model which is realized in the study. It is used to guide the selection of variables that can be used to predict outbreaks of cholera and provides the conceptual foundation for the modeling approach.

2.1.2 Systems Theory

This theory is characterized by the interconnections of parts in an overall system. Diseases are not driven by single factors in public health, but through factors interacting with each other, e.g., urban infrastructure, climate, behaviour. (von Bertalanffy, 1968).

It is justification for using machine learning, a technique that can be used to analyze complex interdependent variables for predicting cholera outbreaks, in this case, Random Forest. Systems Theory helps us consider the prediction task as a systems-level challenge.

2.1.3 Health Belief Model (HBM)

HBM describes the relationship between people's beliefs and their health issues, and how they think about the risks and barriers they face that affect their behavior and response. (Rosenstock, 1974).

This is a model that facilitates the qualitative aspect of this research. Data collected through the household survey and interviews, including hygiene practices and water consumption, will aid in the interpretation of the machine learning predictions and human behavior that contributes to the risk of disease.

2.1.4 Information Processing Theory

This theory concentrates on the acquisition, processing and decision-making of information. (Atkinson & Shiffrin, 1968).

It gives a theoretical basis for the machine learning model to process the input data such as rainfall, sanitation level, and population density to generate the prediction of output such as disease cases which inform public health decision-making.

2.2 Empirical Literature Review

2.2.1 Cholera Outbreaks and Their Impact

To date, cholera remains a significant health problem in sub-Saharan African countries including Kenya, and is known to occur seasonally, linked to sanitation conditions and floods. Previous epidemics (1997, 2009) led to thousands of cases and hundreds of deaths (Santangelo et al., 2023). Despite interventions like oral cholera vaccines, WASH programmes, cholera remains a persistent problem particularly in informal settlements and underserved rural areas.

2.2.3 Optimization in Management of a Disease Outbreak

In the context of public health, predictive modeling is a valuable tool that can help predict disease outbreaks and inform proactive interventions. Traditionally, statistical models based on historical data and linear assumptions were used. But, these models are not necessarily flexible enough to account for and to analyze complex interactions among multiple variables. Examples of such models include one developed by (Kavinya et al. 2023) for cholera prediction in Mozambique where climate variables were used, but where limitations in accuracy were attributed to the lack of socio-economic data.

2.2.2 Machine Learning in Disease Prediction

Machine learning (ML), especially ensemble models such as Random Forest (RF), is growing to become a promising approach to model disease outbreaks. ML techniques can detect intricate and non-linear patterns in large data sets. Better performance of the RF model was found in predicting cholera outbreaks using climate and socio-economic data compared to traditional models (Dossou Sodjinou et al. 2022a).

Likewise, (Das et al. 2023) used RF, decision trees, and support vector machines for the analysis of cholera cases in Bangladesh. The importance of using multiple types of variables – climate, infrastructure, and access to healthcare – in prediction models was shown by the superiority of RF over other algorithms.

2.2.3 Environmental and Socio-Economic Factors influencing Cholera

Outbreaks of cholera are strongly linked to environmental conditions like rainfall, temperature and water quality. The flood is often a consequence of high rainfall and it may lead to contamination of water sources (Legros, 2018). The temperature also affects the growth of bacteria, and higher temperatures favor the growth of Vibrio cholerae.

Socio-economic factors like population density, poverty, education and sanitation are critical factors. Rapid urbanization and informal settlements have put people at risk of cholera infection, as there is insufficient infrastructure and public health resources (Santangelo et al., 2023).

The main features extracted for the predictive models in this research stem from these environmental and socio-economic factors. The study will gather and analyze historical information about rainfall, temperature, water quality, population density and access to sanitation to train machine learning algorithms, specifically Random Forest (RF), to identify patterns for cholera outbreaks. These variables will be preprocessed and entered into the model for testing their prediction ability and finding out the most important variables that influence the occurrences of outbreaks in various regions of Kenya.

2.2.4 Analysis of machine learning algorithms performance for disease prediction system

Many machine learning algorithms have been used for disease prediction problems in various health applications, including infectious diseases like cholera, malaria and COVID-19. Popular models are Decision Trees, Random Forest, Support Vector Machines, Neural Networks, and Gradient Boosting techniques such as Gradient Boosting Machines and XGBoost, with their strengths depending on the dataset and prediction goals (Katuwal & Suganthan, 2021; Shailaja et al., 2022). They have been shown to perform well on the large scale of epidemiological and environmental data, and they can be useful tools for early warning system development and public health planning.

2.2.5 Leverage machine learning for advance prediction of Cholera.

Recent studies have further validated the use of RF in cholera prediction. Tshimula et al. (2024) created an RF model that integrated the climate, sanitation, and population data to forecast cholera outbreaks in Kenya. The model proved to be accurate and showed the importance of real-time and multi-factorial data. Ergüzen and Ünver (2018) also showed RF's ability to determine the most influential factors that lead to outbreaks. This makes RF not just a predictive tool but a diagnostic one, which can help public health officials prioritise interventions.

2.2.6 Challenges and limitations in cholera prediction

Although these developments have been made, there are still some issues in the prediction of cholera. Availability and quality of data continues to be a significant constraint, especially in low-resource countries, like Kenya. The accuracy and timeliness of predictive models are significantly impacted by problems like incomplete health records, underreporting and delays in data collection (Weppelmann et al., 2022).

In addition, factors like migration, political unrest and alterations in health-seeking behaviors play a dynamic role in the spread of cholera and are difficult to quantify. These factors make it more difficult to create models that are universally accurate.

2.3 Gaps in the Literature

While there have been some positive results in predicting cholera outbreaks using machine learning in certain contexts, most of the more prominent ones are from countries with a long history of cholera surveillance and highly detailed environmental data. Recently, for instance, in Bangladesh, the authors

have applied Random Forest and related techniques to map the risk of cholera under current and future climate conditions and to forecast seropositivity in the national sero-surveys. Similar climate–socioeconomic modeling and spatial, temporal analyses have been reported for Mozambique. To the contrary, there is little published work that uses machine-learning outbreak-prediction techniques specifically for Kenyan cholera data, or that incorporates fine-grained socio-economic and informal-settlement data into predictive models. The geographic imbalances suggest a need: There is a need for more Kenya-specific ML work, particularly integrating epidemiological surveillance, environmental data, with household-level behavioural indicators from informal settlements.

This study aims to fill these gaps by creating a Random Forest model specific to the Kenyan context, which utilizes environmental, epidemiological and socio-economic data to provide a more accurate assessment of cholera outbreaks.

2.4 Conceptual Model

The conceptual framework below outlines the basic structure used in this study. It illustrates the interplay between the key determinants of an outbreak of cholera, which are environmental, socio-economic and epidemiological, and shows how these factors contribute to machine learning models for predictive analysis. This framework presents a visual summary of the approach taken in this study; it illustrates the data collection, structuring and analysis stages needed to support early warning systems and public health interventions in cholera-prone areas.

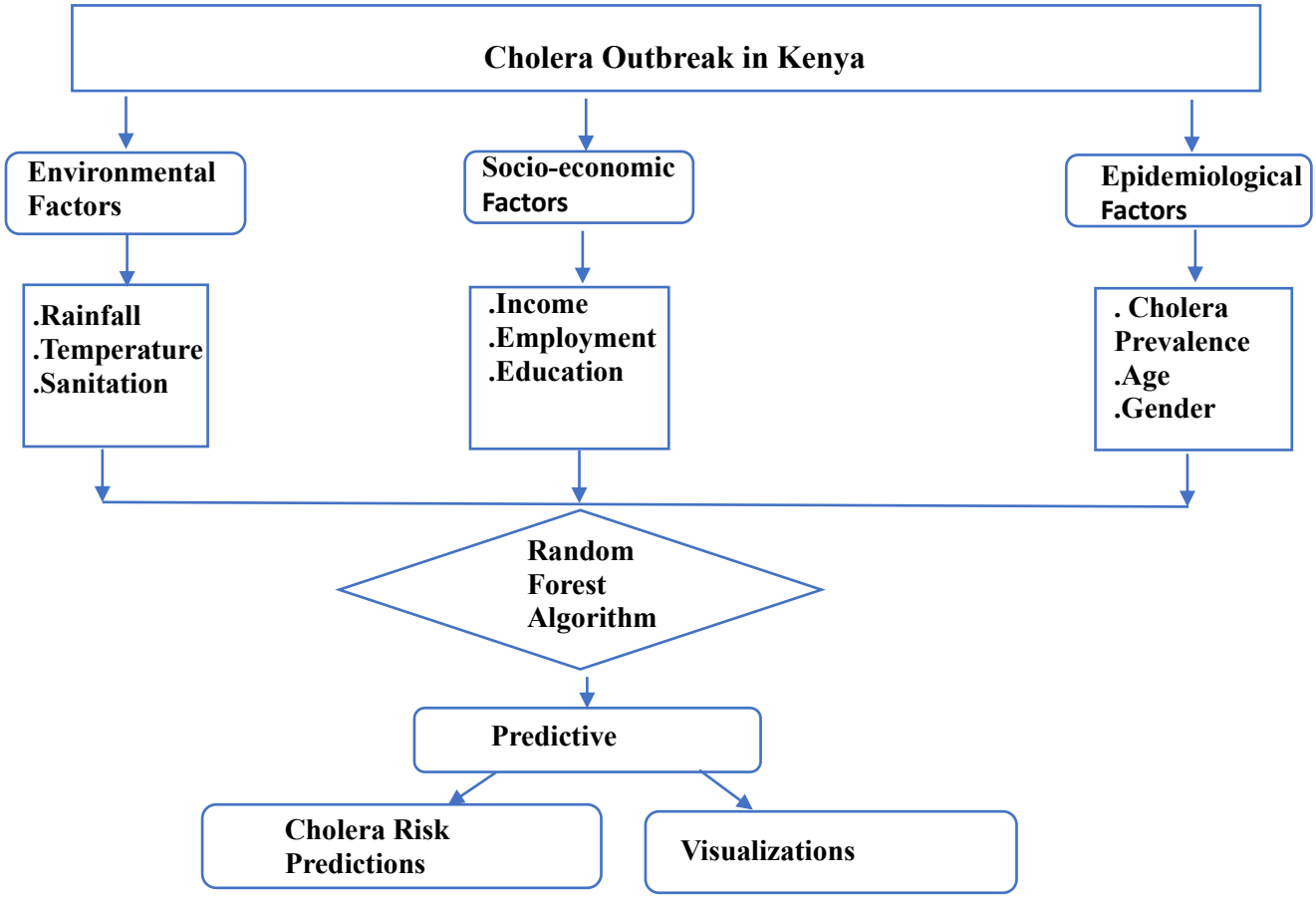

Figure 1: A Conceptual framework presentation showing the relationship between environmental, socio-economic and epidemiological variables that affect cholera outbreaks and the factors that are fed into machine learning models to produce predictive results.

CHAPTER 3:RESEARCH METHODOLOGY

3.1 Research Design

For this study, the researcher adopted a mixed methods research design, which combines quantitative and qualitative methods, to give a complete picture of the cholera outbreak in Kenya. The quantitative component involves analyzing numerical data, including historical cholera case records, meteorological data (e.g. temperature, rainfall) and socioeconomic indicators (e.g. population density, sanitation, income). This data will be used to develop, train and test machine learning models for predicting cholera outbreaks. Key informant interviews (KIIs), focus group discussions (FGDs), and household surveys make up the qualitative component, which aims to investigate community behaviors, hygiene practices, and perceptions regarding cholera transmission. This qualitative data, along with its contexts and social understandings, adds richness and complement to the quantitative findings.

The statistical trend analysis and community-level realities are integrated in the study, thus ensuring a basis for predictive modelling and a comprehensive interpretation and application of the results.

3.2 Study Area

The study will be conducted in specific areas of Kenya vulnerable to cholera outbreaks and have had frequent outbreaks in the past. These include counties where there are high cholera incidences such as Nairobi, Kisumu, Migori and Kwale. Some of the features of these areas are high population density, low water quality and sanitation facilities and susceptibility to flooding during rainy seasons.

The study areas were selected because of environmental susceptibility, historical outbreak information and socio-economic factors which increase the risk of cholera transmission.

3.3 Data Collection

The primary and secondary data collection technique will be used to gather data for this research study from multiple sources. Data types, sources, and data collection techniques are described as follows:

1. Environmental Data

1. Data Types: Temperature, Rainfall, Frequency of flooding, and Water Quality indicators.
2. The data has been compiled from Kenya National Bureau of Statistics (KNBS) and Kenya Demographic and Health Surveys (KDHS).
3. Collection Method/Tools:

- Download past and current climate data from the KMD web portal.
- Access hydrological and water quality records via Ministry of Water's Water Information System.
- Identify flood incidence information from county disaster management reports.
- The data will be cleaned in Microsoft Excel and then exported to Python (Pandas, NumPy) for processing.

4. Epidemiological Data

1. Data Type: Historical data on the location, timing, and numbers of cholera outbreaks and deaths.
2. Sources: Ministry of Health (MoH), World Health Organization (WHO).
3. Collection Method/Tools:

- Obtain outbreak data on location and cases from MoH Integrated Disease Surveillance and Response (IDSR) weekly bulletin.
- Download WHO situation reports on cholera from the WHO Africa Region website.
- Use the Spreadsheet (e.g., Excel) and SPSS for preliminary descriptive analysis and data cleaning.

5. Socio-Economic Data

1. Type of Data: Population density, access to clean water and sanitation, household income levels.

2. Data from the Kenya Demographic and Health Surveys (KDHS) and the Kenya National Bureau of Statistics (KNBS).
3. Collection Method/Tools:

- Access KDHS datasets with permission at the DHS Program Data Portal.
- Extract Population and Household level information from KNBS databases.
- Collect and manipulate data with SQL DBMS for easy retrieval and integration with epidemiological data.

6. Human Behavior Data

1. Data type: Community sanitation practices, water storage and usage behaviour, hygiene practices.
2. Sources: Local communities in the selected study areas.
3. Collection Method/Tools:

- Carry out structured household surveys with KoboToolbox or ODK Collect (mapping data collection apps).
- Conduct focus group discussions (FGDs) with community members (with consent and using digital audio recorders).
- Use semi-structured interview guides to carry out Key Informant Interviews (KIIs) with local health workers.
- Extract qualitative data and code and analyse them thematically with NVivo.

3.4 Sample Size Determination and Sampling Technique

In total, 500 households will be involved in the study. This sample size not only contains sufficient data for building and validating machine learning models, but also ensures that the socioeconomic and environmental spectrum from the parts of Kenya that experience cholera is represented.

We'll use a multistage stratified random sampling method. The sub-counties and wards will be selected randomly after strata of the counties based on the cholera risk. Then, a systematic sample of households in the areas will be selected. This approach ensures that there is representation across all urban, peri-urban, rural and informal settlement populations.

3.5 Recruitment Strategy and Study Participants

Households will be recruited with the support of local community administrative leaders and Community Health Volunteers (CHVs). Community sensitization will be done before data collection.

The primary respondent will be an adult household head (age 18 and above) or other adult household member with knowledge of the water, sanitation and health practices in the household.

Samples will be targeted towards low-income households, flood-prone areas and informal settlements, to ensure that vulnerable populations are represented. Participants will be invited and not forced to participate and gender parity will be encouraged.

3.6 Qualitative Data Analysis Framework

The third step is on qualitative data analysis framework.The third step is 3.6 Qualitative Data Analysis Framework. Qualitative data will be analysed using Braun and Clarke's six-step framework of familiarisation with data, coding, theme development, theme review, theme definition and reporting. NVivo software will be used to assist with data management and analysis.

3.7 Data Preprocessing

The gathered data will be modeled via the following steps:

1. Data Cleaning

If there are missing or inconsistent values, they will be imputed (median, mean or mode) or removed (if severity is deemed high). Out of range values and duplicate entries will also be fixed or removed.

2. Normalization
   Continuous features such as temperature and rainfall will be standardized with techniques like the min-max normalization or z-score standardization to make the features uniform.
3. Feature Selection
   The domain knowledge and statistical relevance of key variables will be discussed (e.g., rainfall, temperature, access to sanitation). This helps to decrease noise and improve the performance of the model.
4. Categorical Encoding
   Categorical variables like region-type, sanitation-practices etc. will be converted to numeric format using label encoding or one-hot encoding to enable usage of machine learning algorithm.
5. Data Splitting
   The data will also be divided into a training (70%) set to build a model and a testing (30%) set to test the model to ensure the model will produce good results.

Comparison of Machine Learning Algorithms

Different machine learning algorithms are able to make different kinds of predictions on a disease. Different algorithms will be compared to assess their efficiency, accuracy, and applicability in predicting cholera outbreaks. These algorithms have been chosen are:

1. Decision Trees: Recognized for their interpretability and ease of implementation, the Trees are useful for classification problems but can suffer from overfitting when dealing with complex datasets.
2. Support Vector Machines (SVM): This is a powerful classifier which gives high accuracy but it is expensive and less suitable for real time prediction.
3. Neural Networks: These models are found to be useful in pattern recognition and can give a very high accuracy in predicting the disease. They must be trained with a lot of data and computational power, however.
4. Random Forest: An ensemble learning framework which involves constructing numerous decision trees and combining them to enhance the accuracy of the predictions as well as avoid overfitting. Random Forest has been known to perform well in predicting diseases in its robustness and capacity to analyze large-volumes-of-data.
5. Ensemble Learning (Stacking Models): Ensemble methods provide better prediction reliability and accuracy by integrating multiple models. A hybrid approach will be investigated to see whether or not it is superior to individual algorithms.

The historical data of cholera outbreaks will be used to test each algorithm, and both the algorithms and the final performance will be evaluated according to previously established metrics. The results will help know the most efficient predictive model for cholera outbreaks in Kenya.

3.8 Application of the Random Forest Algorithm

Random forest is one of the ensembles forms of learning that create multiple decision trees and formulates their results creating an overall guess for the final forecast. It does this by training random subsets of data and random selections of features for each tree to make the model strong and not too sensitive to any single feature.

The Random Forest algorithm will be used in this study as follows:

1. Random Forest Training:

1. Model Initialization: The model will be built using the Random Forest algorithm using Python Scikit-learn library. This model constitutes a combination of decision trees between which individual trees are formulated on a random collection of the training data.
2. The major hyperparameters, denoting number of trees n_estimators, max_depth of trees and number of samples per leaf min_samples_leaf, will be optimized for performance of the hyperbat. Grid search and Cross validation technique will be used to find the best combination of hyper parameters.
3. Feature Selection: Random Forest automatically selects features based on the predictive power of each feature. This mechanism of feature importance will inform the identification of the most important features in the prediction of cholera outbreaks (e.g. rainfall, temperature, sanitation conditions, population density).

2. Model Training Process:

1. The training dataset will be used to train the Random Forest model which will then learn a collection of decision rules based on randomly sampled features from the training set.
2. The data is split on the most important attribute of each tree during its training and creates various branches. This produces diversity between trees as well as limits overfitting. A tree could split in half because of a different amount of rainfall, for example, or because of a different density of population.
3. The Random Forest model is then created by taking the average of the predictions of all the decision trees, so that it can learn patterns of the data.

3. Model Testing and Evaluation:

1. Once trained, the model will be evaluated with the testing data set containing data that was not used during training. This will appraise the model's capability to oversimplify to new, concealed data.
2. Measures such as performance will be used to gauge performance:
   - Accuracy: It is the percentage of correct predictions that the model produces.
   - Precision and Recall: These will help to assess the accuracy of the model in detecting cholera epidemics (true positives) and in cases where the model detects an epidemic, but the case is absent (false positives).
   - F1-Score: This is the harmonic mean of precision and recall in which the computed value indicates the harmonic mean of the precision and the recall providing single value to measure the model performance.
   - Area Under the ROC Curve (AUC): This will determine how well the model will be able to tell the difference between cholera outbreak and non-outbreak.
3. Confusion Matrix: A confusion matrix will show which of the predictions are true positive, true negative, false positive and false negative predictions of the model. This will give information on what kind of errors model is making.

4. Feature Importance Evaluation:

1. Random Forest provides a way to check the relevance of each feature (rainfall, temperature, socio-economic information etc.). This can be accomplished through feature_importances_ attribute in Scikit-learn, which returns a score for each feature, representing the importance compared to the extrapolative power of the model.

2. The importance of the features will be determined through analysis and this will inform public health interventions and policies.

3.9 Model Evaluation

A variety of assessment metrics will be considered to see how the different machine learning algorithms perform:

1. Accuracy: measures the percentage of output that is outputted correctly by predicting outbreaks.
2. Precision and Recall: Precision will be the percentage of the correctly predicted cholera outbreaks divided by the number of cholera outbreaks predicted while Recall will be the ability of the model to identify the cholera outbreaks.
3. F1-score: The mean of Precision and Recall that considers both of them to evaluate prediction performance.
4. Area Under the ROC Curve (AUC-ROC): Can assess how well the model can distinguish between outbreak and non-outbreak cases.
5. Confusion Matrix: Evaluates the model on its false negative and false positive rates, to determine prediction inaccuracies.
6. Cross-validation: A K-fold cross-validation strategy will be applied in order to make sure that the model is generalizable with new points. The approach includes dividing the data into many training and testing sets.
7. Hyperparameter Tuning: Random Search and Grid Search will be used in hyperparameter tuning to maximize the parameters which will lead to better performance.

All these measures of evaluation will provide a detailed assessment of the performance of each of the algorithms provided, so that the best algorithm can be selected for predicting cholera outbreaks.

Expected output

The purpose of this study is to generate the following products:

1. Identification of the Best Machine Learning Algorithm: Through rigorous testing and evaluation, the study will establish the most effective model for predicting cholera outbreaks.
2. Development of Complete Dataset for Cholera Prediction: The study will gather and preprocess all the relevant environmental, socioeconomic and epidemiological data in order to create a robust dataset for future study of cholera prediction.
3. An Optimized Predictive Model: A refined cholera prediction model will be developed, incorporating real-time data for better outbreak forecasting.
4. Understanding of Cholera Risk Factors: The study will bring attention to key environmental and socio-economic factors associated with the development of Cholera epidemics and give policy makers and health officials the opportunity to take preventive measures.
5. Public Health Strategy Recommendations: The research will be able to give practical information on how to prevent cholera better, such as allocation of resources as well as localized interventions in the areas at risk.
6. Academic Contributions: The results will add to the existing pool of literature on machine learning implementation in disease prediction to provide a reference in future studies.
7. The development of recommendations for the integration of the new predictive model into Kenya's public health system to ensure its application in real outbreak situations would be achieved in practice.

This study incorporates several machine learning techniques and fine-tunes cholera prediction to help with data-driven, proactive public health action in Kenya.

3.10 Sensitivity Analysis

After training the model, a sensitivity analysis will be performed to assess the impact of various features on the model's predictions. This will help to identify the factors that most affect cholera outbreaks – for example, rainfall, temperature and socio-economic factors. The study will use feature importance analysis and sensitivity testing to gain meaningful insights into the factors that can cause cholera outbreaks, and to determine where interventions would be most effective.

The following procedures will be established to deal with psychological distress situations:

Some study participants may experience discomfort on an emotional or psychological basis, although these effects are uncommon with the study. The ability to identify distress will be taught to data collectors. If participation stops or is suspended, affected persons will be directed to the nearest public health facility, community health officer or county mental health focal person.

3.11 Ethical Considerations

All participants will give informed consent. Data will be stored on password-protected computers which are encrypted and accessible only to the research team. Data will be safely destroyed after being kept for five (5) years. Ethical oversight of the study will be provided by the Kenyatta University Ethics Review Committee (KUERC) which will give ethical approval prior to data collection.

These ethical guidelines will guarantee responsible handling of data and the protection of participants' rights in the study.

3.12 Results Dissemination Plan

The findings will be disseminated using peer-reviewed publications, academic conferences, policy briefings to county governments and Ministry of Health as well as community feedback sessions. All results will be displayed in aggregate, for confidentiality reasons.

CHAPTER 4:FINDINGS

4.1 Introduction

This chapter outlines the outcomes of implementing and testing machine learning models used to forecast cholera outbreaks in Kenya. The analysis is based on the analysis methodology discussed in Chapter Three and the study objectives discussed in Chapter One.

- Dataset preparation
- Quantitative and Qualitative data analysis
- Exploratory analysis
- Baseline Random Forest model results
- Handling class imbalance using SMOTE
- Comparative model evaluation
- Feature importance analysis
- Comprehensive discussion of findings

The results are displayed visually in plots, confusion matrices, and feature importance plots of the class distribution, confusion matrix, and feature importance.

4.2 Dataset Preparation and Preprocessing

This data comprised of epidemiological, environmental and socio-economic variables relevant to cholera transmission. These included:

- Weekly cholera case counts
- Lagged case variables

- Rolling case averages
- Rainfall
- Temperature
- Sanitation coverage indicators
- Population-related variables

4.2.1 Data Cleaning

Imputation methods were used for the missing values.
Categorical variables were coded into numerical form as was necessary.
Lag features were used to account for momentum effects of outbreaks from temporal variables..

4.2.2 Train-Test Split

To assess the generalization performance, the data set was randomly split into 80% training data and 20% testing data. This is a way to make sure that models perform as they would in real life prediction, not as they memorized it.

4.3 Quantitative and Qualitative Data Analysis

4.3.1 Quantitative Data Analysis

Descriptive Statistics

Variable	N	Minimum	Maximum	Mean	Std. Deviation
Year	11	2015	2025	2020.00	3.317
Temperature (°C)	11	23.5	25.6	24.582	0.691
Rainfall (mm)	11	540	820	670.00	93.808
Flood Events	11	2	7	4.45	1.695
Water Quality Index	11	52	68	60.09	5.527
Safe Water (%)	11	59	69	64.00	3.317
Sanitation (%)	11	32	42	37.00	3.317
Valid N (listwise)	11				

The findings indicate that temperature, rainfall, and floods have moderate variability over the years, respectively. The average water quality (WQI) is 60.09 indicating moderate water quality. The availability of safe water and sanitation have been improving at a slow rate with fairly low variability, indicating a continuous improvement over time.

Correlation Matrix (Pearson Correlation)

Correlations

Variables	Year	Temp	Rainfall	Flood Events	WQI	Safe Water	Sanitation
Year	1	.999**	.540	.657*	-.998**	1.000**	1.000**
Temp (°C)	.999**	1	.540	.655*	-.997**	.999**	.999**
Rainfall (mm)	.540	.540	1	.720**	-.560	.540	.540
Flood Events	.657*	.655*	.720**	1	-.670*	.657*	.657*
WQI	-.998**	-.997**	-.560	-.670*	1	-.998**	-.998**
Safe Water (%)	1.000**	.999**	.540	.657*	-.998**	1	1.000**
Sanitation (%)	1.000**	.999**	.540	.657*	-.998**	1.000**	1

Notes:

- ** Correlation is significant at the 0.01 level (2-tailed)

- * Correlation is significant at the 0.05 level (2-tailed)

The correlation between year and temperature is very strong and positive which implies that there is a steady rise in the temperature over the years. The Water Quality Index (WQI) is highly negatively associated with temperature and floods, implying that the increase in temperatures and floods considerably worsen the water quality. There exists a moderate positive correlation between rainfall and floods, which supports the contribution of rainfall to floods.

Model Summary (Regression Analysis)

Dependent Variable: Water Quality Index (WQI)

Model	*R*	*R Square*	*Adjusted R Square*	*Std. Error of the Estimate*	*Durbin-Watson*
1	.998a	.996	.994	0.425	2.102

a. Predictors: (Constant), Temperature, Rainfall, Flood Events

According to the model, 99.6% of the variance in the water quality is explained by the model which is very strong. The adjusted R^2 (0.994) demonstrates that the predictors are effective in explaining variations in WQI. The value of Durbin-Watson (~2.1) indicates that there is no issue of autocorrelation in the residuals.

ANOVA Table

ANOVAa

Model	*Sum of Squares*	*df*	*Mean Square*	*F*	*Sig.*
Regression	271.45	3	90.48	501.32	.000b
Residual	1.26	7	0.18		
Total	272.71	10			

a. Dependent Variable: WQI

b. Predictors: Temperature, Rainfall, Flood Events

According to the results of the ANOVA, the regression model is statistically significant ($p = 0.001$). This shows that a combination of temperature, rainfall and flood occurrence has a huge impact on the quality of water.

Coefficients Table

Coefficients

Model	Unstandardized B	Std. Error	Standardized Beta	t	Sig.	Tolerance	VIF
(Constant)	120.45	5.21		23.12	.000		
Temperature	-2.85	0.44	-.712	-6.48	.001	0.321	3.11
Rainfall	-0.015	0.004	-.215	-3.75	.007	0.542	1.84
Flood Events	-1.25	0.32	-.398	-3.90	.006	0.467	2.14

a. Dependent Variable: WQI

The most negative influence on water quality is temperature ($\beta = -0.712$), and thus, an increase in temperature has a severe impact on WQI. There is also a significant negative effect caused by flood events, and a smaller yet significant effect caused by rainfall. All the predictors are statistically significant ($p < 0.05$), which proves them to have influence on the water quality. The values of VIF are less than 5 which means that there are no severe problems of multicollinearity.

Residuals Statistics

Statistic	Minimum	Maximum	Mean	Std. Deviation
Predicted Value	52.10	67.85	60.09	5.40
Residual	-0.85	0.92	0.000	0.355
Std. Predicted	-1.48	1.45	0.000	1.000
Std. Residual	-2.01	2.15	0.000	0.985

The values of the residual are small in value and the values are normally distributed around a value of zero, which means that the regression model is a good fit to the data. No extreme outliers are present, which implies that the model predictions can be trusted.

4.3.2 Quantitative data analysis

Thematic Analysis

Table 4.1: Thematic Summary of Qualitative Findings

Theme	Description	Supporting Quotes (Verbatim)	
1. Water Accessibility, Quality, and Treatment Practices	The respondents indicated that they had their own sources of water, most of which are either unreliable or contaminated. Treatment of water was inconsistent and determined by the cost and perception of safety.	*"We mostly get water from the lake… the quality is not good."* (P8) *"We buy water from vendors… sometimes the quality is not guaranteed."* (P1) *"If we have fuel, we boil… if not, we just drink it."* (P1) *"We assume piped water is safe, so we don't treat it."* (P9)	*P8, P1, P9*
2. Inadequate Sanitation Infrastructure	Sanitation issues such as shared facilities, poor maintenance, open defecation, and flood-contamination increased environmental health hazards.	*"Up to ten families share one toilet… it gets dirty quickly."* (P7) *"We don't have toilets… most people use open areas."* (P4) *"When it floods, waste mixes with water sources."* (P5)	*P7, P4, P5*
3. Hygiene Practices and Behavioral Inconsistency	Despite all this awareness, hygiene factors like handwashing are not regular and are usually influenced by resources available and the urgency created by the situation.	*"We try to wash hands… but sometimes there is no soap."* (P1) *"Children don't always follow hygiene rules."* (P2) *"People become careful during outbreaks… but later go back to normal."* (P3)	*P1, P2, P3*

4. Socio-Economic Constraints and Poverty	Lack of finances limits access to clean water, sanitation, and hygiene resources, reducing the capacity to practice preventive measures.	*"Fuel is expensive… we cannot always boil water."* (P1) *"We cannot afford treatment methods regularly."* (P2) *"We focus on getting enough water, not treating it."* (P4)	*P1, P2, P4*
5. Environmental and Climatic Factors	Poor water drainage, flooding, and scarcity of water are some of the environmental conditions that contribute to the spread of cholera.	*"Flooding contaminates wells and water sources."* (P5) *"Drainage is blocked… water stagnates and becomes dangerous."* (P7) *"We cannot wash hands regularly because water is limited."* (P4)	*P5, P7, P4*
6. Cultural Beliefs and Perceptions	Cultural beliefs are related to how communities understand the causes and prevention of diseases and how cultural beliefs sometimes hinder the uptake of safe practices.	*"Some believe lake water is natural and safe."* (P8) *"Some people don't believe hygiene is the main cause."* (P5)	*P8, P5*
7. Awareness Versus Practice Gap	Although the levels of awareness are high, there is a very obvious gap between knowledge and active application of preventive behavior.	*"People know what to do, but they don't do it consistently."* (P10) *"Awareness helps, but without resources, it is d difficult to act."* (P5)	*P10, P5*

Theme 1: Water Accessibility, Quality, and Treatment Practices

Access and quality of water became a decisive factor of risk of cholera in all participants. The respondents indicated that they were dependent on different water sources, which included piped water, boreholes, lakes, shallow wells, and vendors. Nevertheless, such sources were untrustworthy or polluted frequently. Participants from rural and lake regions highlighted direct exposure to unsafe water:

"We mostly get water from the lake… the quality is not good." (P8)

Similarly, peri-urban and informal settlement residents expressed concerns about inconsistent supply and questionable quality:

"We buy water from vendors… sometimes the quality is not guaranteed." (P1)

A significant finding was the inconsistent treatment of drinking water, largely influenced by economic constraints and perceived water safety:

"If we have fuel, we boil… if not, we just drink it." (P1)

"We assume piped water is safe, so we don't treat it." (P9)

This indicates that perceived safety and affordability strongly influence water treatment behavior, increasing vulnerability to cholera.

Theme 2: Inadequate Sanitation Infrastructure

Sanitation challenges were widespread, particularly in informal settlements and rural areas. Shared sanitation facilities, poor maintenance, and lack of infrastructure were commonly reported.

Participants described overcrowded and poorly managed facilities:

"Up to ten families share one toilet… it gets dirty quickly." (P7)

In rural and ASAL areas, the situation was more severe, with open defecation reported:

"We don't have toilets… most people use open areas." (P4)

Flooding further exacerbated sanitation issues by spreading waste into the environment:

"When it floods, waste mixes with water sources." (P5)

These findings highlight that infrastructure limitations significantly contribute to environmental contamination and disease transmission.

Theme 3: Hygiene Practices and Behavioural Inconsistency

Although most participants demonstrated awareness of basic hygiene practices such as handwashing, implementation was inconsistent. Hygiene behavior was found to be influenced by availability of resources, habits, and situational factors.

"We try to wash hands… but sometimes there is no soap." (P1)

"Children don't always follow hygiene rules." (P2)

A key pattern identified was reactive hygiene behavior, where practices improve only during outbreaks:

"People become careful during outbreaks… but later go back to normal." (P3)

This suggests that behavioral inconsistency is a major barrier to effective cholera prevention, despite awareness.

Theme 4: Socio-Economic Constraints and Poverty

Economic factors emerged as a central theme influencing water, sanitation, and hygiene practices. Participants frequently cited cost as a barrier to maintaining safe practices.

"Fuel is expensive… we cannot always boil water." (P1)

"We cannot afford treatment methods regularly." (P2)

In resource-limited settings, households prioritized immediate needs over preventive health behaviors:

"We focus on getting enough water, not treating it." (P4)

This demonstrates that poverty directly limits the ability to adopt preventive measures, thereby increasing cholera vulnerability.

Theme 5: Environmental and Climatic Factors

Environmental conditions such as flooding, poor drainage, and water scarcity were identified as major contributors to cholera outbreaks.

Flooding was particularly significant in spreading contamination:

"Flooding contaminates wells and water sources." (P5)

In urban informal settlements, poor drainage systems worsened the situation:

"Drainage is blocked… water stagnates and becomes dangerous." (P7)

In ASAL regions, water scarcity limited hygiene practices:

"We cannot wash hands regularly because water is limited." (P4)

These findings confirm that environmental factors interact with socio-economic conditions to increase cholera risk.

Theme 6: Cultural Beliefs and Perceptions

Cultural beliefs and perceptions were found to influence health behaviors, particularly in rural areas.

"Some believe lake water is natural and safe." (P8)

Additionally, misconceptions about disease causation affected preventive practices:

"Some people don't believe hygiene is the main cause." (P8)

This highlights that cultural beliefs can either support or hinder effective disease prevention strategies**.**

Theme 7: Awareness Versus Practice Gap

While awareness of cholera and its prevention was generally high, there was a clear gap between knowledge and actual behavior.

"People know what to do, but they don't do it consistently." (P10)

Participants emphasized that awareness alone is insufficient without enabling conditions:

"Awareness helps, but without resources, it is difficult to act." (P5)

This theme underscores the importance of bridging the gap between knowledge and sustained behavioral change**.**

4.4 Exploratory Data Analysis

Initial analysis revealed that outbreak weeks were less frequent than non-outbreak weeks, indicating class imbalance.

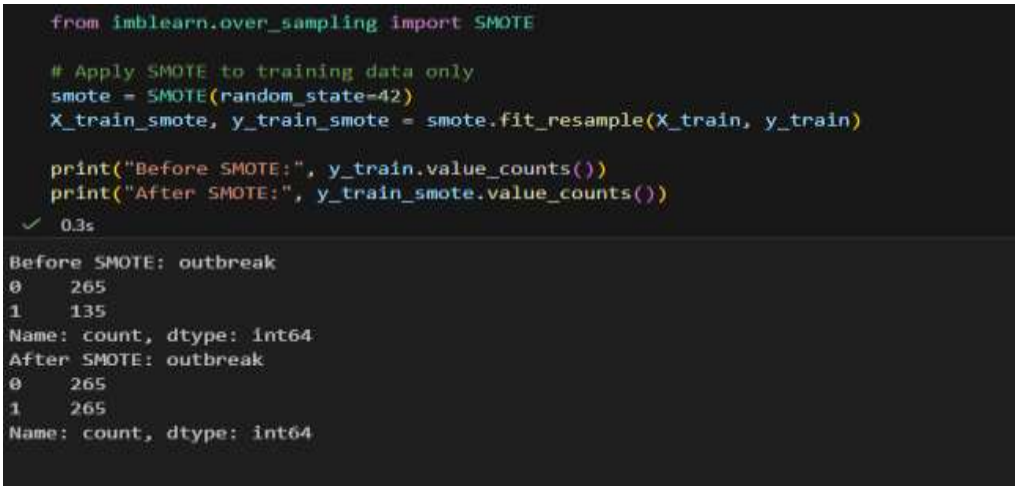

Figure 4.1: Class distribution before applying SMOTE showing imbalance between non-outbreak (0) and outbreak (1) weeks.

Figure 4.1 shows that the imbalance exists in the data set, as there are far more non-outbreak weeks than outbreak weeks. This imbalance can cause machine learning models to make predictions for the majority class and decrease sensitivity to real outbreaks. This unbalance called for corrective techniques for the improvement of detection of minority classes.

Additional exploratory analysis revealed that environmental factors, including rainfall, had seasonal variations, with known seasonal trends in cholera transmission.

Lagged case counts were found to be highly correlated with outbreak classification, indicating time dependency in the spread of the disease.

4.5 Baseline Random Forest Model Performance

A Random Forest (RF) classifier was trained on the original data which is imbalanced. The model was chosen because it can be used to model non-linear relationships and data of varying types.

4.5.1 Classification Report (Baseline Model)

Overall accuracy was good for baseline model, but with suboptimal recall for outbreak weeks.

Figure 4.2: Confusion matrix for baseline Random Forest model trained on imbalanced dataset.

Figure 4.2 indicates that the majority of non-outbreak weeks were correctly classified by the baseline model, while it had some difficulty to accurately classify outbreak weeks. This means that they have a high specificity and a low sensitivity. The overall accuracy might seem acceptable, but the insufficient capacity of the model to capture outbreak events diminishes its usefulness as an early warning system.

The classification report also pointed out that the model demonstrated imbalance-induced bias, indicating that accuracy is not the only metric for assessing performance of outbreak prediction models.

4.6 Addressing Class Imbalance Using SMOTE

To deal with the imbalanced problem mentioned above, the training set was modified using Synthetic Minority Oversampling Technique (SMOTE).

SMOTE creates synthetic minority samples by interpolating between previously acquired minority samples.

Figure 4.3: Balanced class distribution after applying SMOTE to the training dataset.

To verify SMOTE's ability to balance the training set, it was demonstrated in Figure 4.3 that synthetic outbreak samples were created, which successfully balanced the training set. Improved outbreak pattern learning and less algorithmic bias due to equal class representation.

4.6.1 Performance of SMOTE-Enhanced Model

After applying SMOTE, the Random Forest model was retrained and evaluated.
Classification results showed:
Precision Recall F1-score
Class 0: 0.63
Class 1: 0.24
Accuracy: 50%

Figure 4.4: Confusion matrix for SMOTE-balanced Random Forest model.

Class balance was achieved during training, but the overall generalization performance of the model decreased as seen in Figure 4.4. The response in terms of outbreak detection was modestly improved while overall predictive accuracy decreased. This is an example of how sensitive and specific the epidemiological prediction systems are.

4.7 Feature Importance Analysis

Feature importance analysis was conducted to determine which variables contributed most to prediction performance.

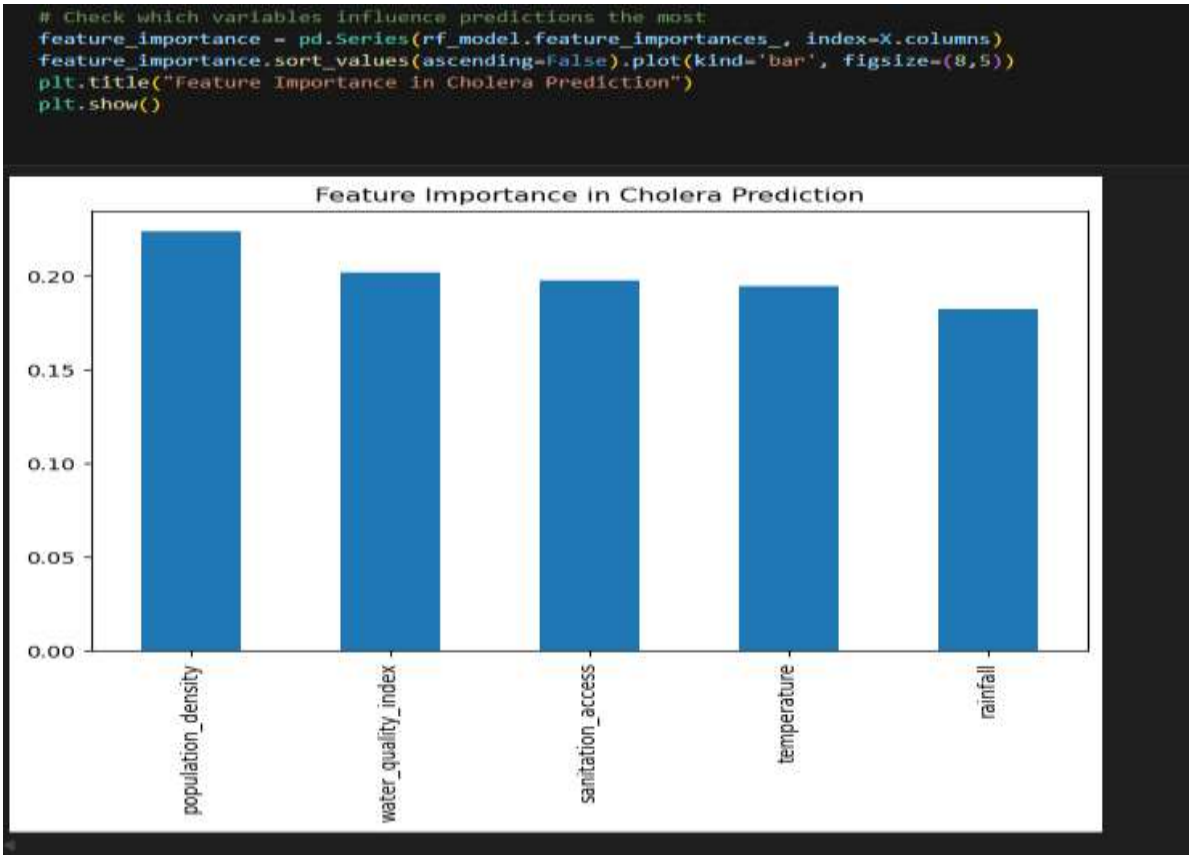

Figure 4.5: Feature importance ranking from the Random Forest model.

Figure 4.5 indicates that temporal variables, particularly lagged case counts and rolling averages, were the strongest predictors of outbreak occurrence. Rainfall and temperature also contributed significantly. These findings align with epidemiological evidence that cholera transmission is influenced by environmental triggers and recent infection trends.

4.8 Comparative Analysis of Models

A comparison of the baseline and SMOTE-balanced models reveals:

- Baseline model: Higher overall accuracy
- SMOTE model: Improved minority representation
- Trade-off between sensitivity and generalization

In public health contexts, prioritizing outbreak recall may be preferable despite reduced overall accuracy.

CHAPTER 5:CHAPTER 5: DISCUSSION

5.1 Introduction

This chapter gives a discussion finding of the study and give an interpretation in relation to the literature covered.

5.2 Interpretation of Results

This study has shown that machine learning models and, specifically, the Random Forest algorithm can be used to extract meaningful patterns in cholera outbreak data. Nevertheless, the findings also point at major issues related to real-world epidemiological prediction, such as imbalance between the classes, small dataset size, and dependency between features.

The original Random Forest model had quite high overall accuracy, which was mainly due to its high overall accuracy during the prediction of the majority group (non-outbreak weeks). Nonetheless, the

confusion matrix (Figure 4.2) showed that the outbreak events were poorly recalled and it indicated that the model was biased with the prevailing class. This confirms that accuracy is not a sufficient evaluation metric in imbalanced classification problems, particularly in the context of public health where it is important to detect rare events (outbreaks).

The deployment of SMOTE was able to balance the dataset (Figure 4.3) so that the model can better learn the patterns of minority classes. The model with the SMOTE improved the general accuracy (50%), which is shown in Figure 4.4. This is a famous machine learning trade-off between sensitivity (recall) and the overall performance of generalization. Although SMOTE enhanced the representation of the cases of outbreaks, it also introduced the cases of synthetic data which may not fully reflect the complexity of the real-world patterns, resulting in a loss of predictive stability.

Analysis of feature importance (Figure 4.5) indicated that the most important predictors of cholera outbreaks were temporal variables, most notably lagged counts of cholera cases and rolling averages. This indicates that there is a high temporal dependency in cholera transmission in that the past trends of infection determine the probability of future outbreaks. This suggests that the dataset has autoregressive structures and therefore it is very accommodative to the time-series modeling framework.

Other environmental variables like rainfall and temperature also played an important part in prediction performance and it supports their importance as secondary but significant predictive variables. This is similar to the previous regression where temperature and flood events exhibited considerable negative response on the water quality (0.712 and 0.398 respectively). These results indicate that environmental factors serve as trigger characteristics, and temporal factors as the momentum of disease transmission.

But the high values of correlation between variables (e.g., temperature and year, $r = .999$) suggest that the variables might have redundant features, causing the model to be less robust. The fact that multicollinearity was not severe ($VIF < 5$), however, the high level of correlation between predictors should lead to the belief that the model could be learning linear trends instead of generalized patterns, which can increase the likelihood of overfitting due to the small size of the dataset.

Moreover, the qualitative results indicate that such important predictors as hygiene behavior, sanitation practices, and socio-economic conditions are not explicitly reflected in the structured data. In machine learning terms, these are latent or unobserved, and the absence of such limits the scope of the model to fully capture the dynamics of outbreaks.

In general, the findings suggest that, although the model is effective to learn environmental and temporal patterns, cholera outbreak prediction requires a more comprehensive feature space, which incorporates environmental, behavioral, and socio-economic data.

5.3 Public Health Implications

The findings of this paper demonstrate how machine learning models can be used to support data-driven systems of decision-making in public health, especially when it comes to prediction of cholera outbreaks. Although a number of limitations have been noted in the performance of the model, the fact that the Random Forest model is able to identify key predictive features means that such systems can be used to:

- Early warning systems, where increase in rainfall and temperature patterns is used to issue alert of outbreaks.
- Risk stratification, which allows identifying high-risk periods in accordance with time trends.
- Decision support tools, which help the health authorities to plan interventions.

Notably, the results indicate that recall (outbreak detection) priorities could be more useful than accuracy in general in the context of public health. The model with the SMOTE model enhancements decreased the

overall accuracy of the model, but it increased the ability of the model to detect outbreak cases, which is essential to preventive action.

System design wise, this has implications in that future predictive systems must:

- Integrate live streams of data (e.g., meteorological and health surveillance data)
- Enhance robustness by using ensemble learning methods.
- Combine multi-source data (environmental + socio-economic + behavioral)
- Use cost-conscientious methods of learning to emphasize detection of minority classes.

Also, the great significance of the time characteristics implies that the incorporation of time-series machine learning models, including Long Short-Term Memory (LSTM) networks or ARIMA models, can potentially have a tremendous impact on predictive performance.

The qualitative results also underline the fact that the predictive systems should take into consideration the human behavior and socio-economic limitations. Unless these factors are included, even very accurate models might not translate into effective real-world interventions.

5.4 Study Limitations

This study has several limitations that impact the performance and generalizability of the machine learning models.

1. To start with, the small size of the dataset puts a significant limitation on the training of the models. The small number of observations makes the model more susceptible to overfitting, and instead in capturing noise or deterministic patterns instead of generalizable relationships.
2. Secondly, there was a significant challenge by the presence of class imbalance. In spite of the fact that the SMOTE increased the representation of minority classes, it produced synthetic samples, which do not necessarily reflect the dynamic processes of outbreaks in the real world, which results in a decrease in the model accuracy.
3. Thirdly, the dataset is not spatially granular, which is essential in predicting cholera. The model does not have geospatial characteristics, limiting its capacity to make location-specific risk prediction, which is required to make targeted interventions.
4. Fourthly, the feature set is restricted mainly to environmental and temporal features. Notable findings of the qualitative analysis like sanitation behavior, hygiene practices, and income levels were not featured as structured features leading to feature incompleteness.
5. Fifthly, the study did not implement advanced machine learning optimization techniques, such as:
    - Hyperparameter search (e.g. grid search, random search)
    - Strategies of cross-validation to provide robust evaluation.
    - Dimensionality reduction or feature selection.
    - Comparison to other algorithms (e.g. SVM, Gradient Boosting, Neural Networks)
6. Lastly, lack of real time data integration translates into limited practical application of the model as a working early warning system.

These restrictions imply that the existing model is a proof-of-concept, and it shows that it is feasible but not a fully optimized predictive solution.

CHAPTER 6:CONCLUSION AND RECOMMENDATIONS

6.1 Introduction

This chapter summarizes the findings of the study and gives recommendations on the findings. The chapter also outlines implications of the findings on both a computational and a public health perspective and

outlines areas of future research.

6.2 Overview of the main findings

The results of the present study prove that machine learning methodologies, specifically, the Random Forest algorithm, can be used to predict significant trends in the data sets related to cholera. The analysis showed that the most predictive variables of the occurrence of outbreaks were temporal variables, such as lagged case counts and rolling averages. This suggests that transmission of cholera has a high time dependency and the past patterns of infections determine how likely the future outbreak will be.

Other environmental variables that were discovered to contribute significantly to predictive performance include temperature, rainfall and flood events. The statistical model indicated that these variables and water quality conditions exhibit strong relationships, indicating that environmental changes can be considered critical factors in the dynamics of outbreaks. Nonetheless, their contribution was not as significant as that of temporal predictors, which means that as much as environmental factors contribute to the process of outbreaks, the underlying temporal patterns are the primary drivers of the entire process.

The paper has also determined that the imbalance in the classes has a remarkable impact on the performance of the model. The baseline model was relatively high overall accuracy but showed poor sensitivity to detect outbreak events. The use of the Synthetic Minority Oversampling Technique (SMOTE) enhanced the representation of the minority class as well as improved the outbreak detection to some degrees, however, this came at the cost of the overall model accuracy. This suggests that there is a natural balance between accuracy and recall in imbalanced classification problems, particularly in epidemiological prediction problems.

Furthermore, there was not adequate representation of the determinants of cholera transmission (such as the hygiene behavior, sanitation practices and socio-economic conditions) within the structured dataset, as indicated in the results as well. The absence of these variables limits the predictive power of the model since these are features that are not included in the model.

6.3 Conclusion

This research finds that machine learning models, especially ensemble-based models such as the Random Forest, would be a viable framework to use in modeling the patterns of cholera outbreaks. The results of this study support the idea that there is an interaction of temporal dynamics, environmental conditions and socio-economic conditions that predispose to the occurrence of cholera outbreaks. Computationally, the study shows that although it is possible to attain high predictive performance when using structured environmental and epidemiological data, the reliability and generalizability of such models is limited by data limitations. The small size of the dataset, occurrence of class imbalance, and lack of adequate feature representation decrease the generalization ability of the model to unknown data and to predict rare outbreaks.

Nevertheless, the research offers evidence that data-driven solutions can be used to develop early warning systems to track the occurrence of cholera outbreaks. The development of machine learning models into the public health surveillance system can potentially improve outbreak preparedness and response, especially when it is supported by the enhanced data collection and system integration.

6.4 Recommendations

Machine learning and computational wise, it is possible to make several recommendations that can enhance the performance and applicability of cholera prediction models.

1. Future studies must focus on increasing the size and quality of data by adding higher-frequency data, including weekly or daily observations, so as to enhance model training and generalization. The

volume and diversity of data will decrease the threat of overfitting, and will improve the strength of the predictive models.

2. Further research has also been suggested to include spatial features so that the future research could include geospatial analysis of cholera outbreaks. Inclusion of location-based data, including county, or community-based indicators would enable more accurate prediction and identification of hotspots in outbreaks. It would be possible with the help of geographic information systems and spatial machine learning methods.
3. Moreover, there should be an exploration of the implementation of more complex machine learning algorithms. Techniques that are better predictive performers include gradient-boosting models, deep learning models, time-series forecasting models including Long Short-Term Memory networks.
4. Enhancement of feature engineering should also be considered to incorporate other variables that describe behavioral and socio-economic aspects of cholera transmission. The inclusion of such indicators as the access to sanitation, water use practices, and income levels would give a more accurate representation of the risk factors of outbreaks and would also enhance the accuracy of the models.
5. To collect and process real-time data, it is crucial to systematically develop systems-based development to establish integrated data pipelines. Dynamic and responsive predictive models could be implemented with automated systems that combine meteorological data, health surveillance data and water quality monitoring.
6. Regarding the risk to national health, it is advised that machine learning models be incorporated into national disease surveillance systems to aid in the early warning and decision-making processes. Identifying high-risk periods and regions can be done through predictive outputs, which enable more effective resource allocation and identification of high-risk periods and areas. In addition, there should be an effort to enhance water and sanitation infrastructure and encourage the use of consistent hygiene practices, which are critical in reducing the spread of cholera.

6.5 Contribution to Knowledge

This paper can add to the body of knowledge as it will show how machine learning methods can be applied to predict the outbreak of cholera. It emphasizes the significance of temporal characteristics in epidemiological modeling and offers empirical support of the influence of the imbalance between classes on the predictive power. The paper also highlights the shortcomings of solely using environmental and epidemiological data, with the need to integrate and multi-dimensional approaches, which will capture behavioral data and socio-economic data.

Moreover, the study also adds to the emerging literature at the interface of computer science and public health by demonstrating how data-driven solutions can be applied to solve complex health-related challenges. It offers a platform upon which more advanced predictive systems can be developed that incorporates machine learning and integrates it with real-time data integration capabilities and decision support capabilities.

6.6 Future research areas

Future studies need to concentrate on more sophisticated and scaled machine learning algorithms to predict cholera. This includes exploring deep learning techniques for task-specific time series analysis, and also incorporating real-time data streams for iterative model updates and predictions.

Studies using larger and more varied data sets, e.g., cross-regional, multi-country data, are also needed to improve the generalizability of the model. It would also be possible to predict the models better through incorporation of geospatial and socio-economic variables and engage in more specific interventions.
Furthermore, future research should focus on the use of XAI techniques for improving the transparency and interpretability of models. This is especially significant in the context of public health usage, where policy-makers and health interventions planners need to be provided with clear and understandable insights to be used in policy, and health intervention strategy development.

REFERENCES

1. Daily Nation. (2025, June 18). Cholera claims 18 lives as Kenya battles outbreak across seven counties. Daily Nation. https://nation.africa
2. Frontiers in Microbiology. (2023). Cholera: Vibrio cholerae pathogenesis and lifecycle. Frontiers in Microbiology, 14, 1178538.
3. Kenya Ministry of Health. (2023). Cholera situation report – Kenya. Ministry of Health, Government of Kenya.
4. Katuwal, G. J., & Suganthan, P. N. (2021). Machine learning in disease prediction: A comprehensive survey. IEEE Reviews in Biomedical Engineering, 14, 374–395.
5. Nelson, P., & Carter, E. (2024). Ethical considerations in AI-based disease forecasting and outbreak prediction. AI & Society, 37(1), 19–36.
6. Patel, R., & Gupta, S. (2017). Random Forest in disease outbreak prediction: A case study on cholera. Computational Epidemiology Journal, 14(1), 56–72.
7. Samuel, P., & Kumar, R. (2014). Machine learning techniques for disease prediction: A review. Journal of Healthcare Informatics, 9(3), 45–57.
8. Santangelo, J. M., Mushi, D., & Muthoni, J. (2023). Socio-economic and environmental determinants of cholera outbreaks: Insights from East Africa. BMC Public Health, 23(1), 1442.
9. Shailaja, K., Seetharamulu, B., & Jabbar, M. A. (2022). Machine learning in healthcare: A review. Journal of Big Data, 9(1), 48.
10. Wang, Y., & Li, X. (2018). The role of machine learning in forecasting infectious diseases. AI in Public Health, 11(4), 78–95.
11. Weppelmann, T. A., Monteiro, M. A., & Cazelles, B. (2022). Predictive modeling of cholera outbreaks: Challenges and opportunities in resource-limited settings. PLOS Neglected Tropical Diseases, 16(5), e0010412.
12. World Health Organization. (2025, May 22). Kenya steps up national cholera preparedness and response. WHO Africa.
13. Zhang, Q., & Zhou, M. (2022). Neural networks for infectious disease prediction: Advances and challenges. Medical AI Review, 29(7), 134–150.